Data Min Knowl Disc (2015) 29:1211–1232
DOI 10.1007/s10618-014-0399-1

MASSEXODUS: modeling evolving networks in harsh environments

Saket Navlakha · Christos Faloutsos · Ziv Bar-Joseph

Received: 26 July 2014 / Accepted: 18 December 2014 / Published online: 7 January 2015

Abstract Consider networks in harsh environments, where nodes may be lost due to failure, attack, or infection—how is the topology affected by such events? Can we mimic and measure the effect? We propose a new generative model of network evolution in dynamic and harsh environments. Our model can reproduce the range of topologies observed across known robust and fragile biological networks, as well as several additional transport, communication, and social networks. We also develop a new optimization measure to evaluate robustness based on preserving high connectivity following random or adversarial bursty node loss. Using this measure, we evaluate the robustness of several real-world networks and propose a new distributed algorithm to construct secure networks operating within malicious environments.

Keywords Graph models · Network robustness · Biological fragility

Responsible editor: Joao Gama, Indre Zliobaite and Alipio Jorge.

S. Navlakha's study began during a post-doc at Carnegie Mellon University.

S. Navlakha (✉)
Center for Integrative Biology, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
e-mail: navlakha@salk.edu

C. Faloutsos
Machine Learning Department, Computer Science Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Z. Bar-Joseph
Machine Learning Department, Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

1 Introduction

Some networks operate under harsh environments, with rampant node loss (e.g. a communication network in a battlefield or a sensor network monitoring an active volcano); while other networks operate under milder environments (e.g. a phone network with customers being occasionally lured away by competitors). If we are shown a static snapshot of a network, can we predict the harshness of the environment to understand how node growth and node loss processes together shaped topology? Can we model and extrapolate these processes to design robust networks?

Prior work has largely studied network robustness assuming single node loss—that is, a node is chosen randomly or adversarially, and it is removed, along with all its edges (Albert et al. 2000). We propose to use *bursty* node loss: not only is the original node removed, but several of its neighboring nodes, and so on recursively, the extent of which is governed by a single parameter, β. The value of β measures the harshness of the environment by indicating how likely a node will be lost in this environment and how contagious this loss will be to neighboring nodes. We show that β correlates well with the harshness of the environment of biological protein interaction networks, and we further develop the MassExodus graph generation model, which can closely mimic the topology of additional real-world graphs.

Most generative models of network evolution assume monotonic growth of nodes over time (Barabasi and Albert 1999; Broder et al. 2000; Fabrikant et al. 2003; Vazquez et al. 2003; Leskovec et al. 2005a, b; Siganos et al. 2006; Leskovec et al. 2008; Akoglu and Faloutsos 2009; Chakrabarti and Faloutsos 2012); however, there are several cases where nodes are lost and in a bursty fashion: For example, infected machines on the Internet can spread viruses or malware to neighboring machines, potentially affecting hundreds of thousands of machines and costing billions to fix (Moore et al. 2002, 2003). On the power grid, nodes can fail when demand exceeds individual node capacity; such failures forces redistribution of load to neighboring nodes, potentially triggering widespread blackouts that affect millions of customers (Albert et al. 2004; Sole et al. 2008). In mobile or sensor networks, devices can collectively malfunction due to battery issues (Carle and Simplot-Ryl 2004), lack of solar power (Alippi and Galperti 2008), or other harsh environmental conditions. Even in social networks, a user may leave the network to join a rival and in doing so may influence some of his friends to leave, as well (Wu et al. 2013; Cho 2013). These instances suggest that bursty failures, coinciding with regular growth, constitute a closer model of reality.

Biological networks present an interesting case to study how evolution has rewired topology to optimize function in the face of bursty node failures, mutations, and noise. In molecular interaction networks, protein failures (Gidalevitz et al. 2011), gene mutations (Gu et al. 2003; Kitano 2004), and propagating environmental and signaling noise (Newman et al. 2006) all affect how collections of molecules or cells process information and coordinate responses. In these networks, gene loss events can be correlated [i.e. two genes are either always present together in the genome, or both absent (Valencia and Pazos 2003)], which also suggests that node loss effects can propagate through the network in evolutionary time. Prior work suggests that the topology of these networks has adapted to minimize the impact of propagating environmental noise on function (Navlakha et al. 2014), yet environmental harshness is typically

not considered within evolutionary models. Thus, the wide spectrum of topologies observed within real biological networks (from highly connected and redundant to sparsely connected and distributed) cannot easily be explained by current generative models for biological networks (Vazquez et al. 2003; Middendorf et al. 2005).

In addition to modeling the harshness of network environments, there is also a need to *build* networks that are optimal for different environments. What is a good measure of "optimality"? While building cliques may not be practical due to limited resources or budget, high connectivity was previously suggested to lead to higher fault tolerance (Albert et al. 2000; Schneider et al. 2011; De Domenico et al. 2014; Chan et al. 2014). However, evidence from real-world networks suggests that such high connectivity is not necessarily more robust. For example, analyses of ecosystems have concluded that population stability sharply transitions from overall stability to instability as the number and strength of interactions amongst species increases (Haldane and May 2011), and similar observations hold for dependencies within financial transaction networks (Acemoglu et al. 2013). Terrorist contact networks are also markedly sparse, which may be a deliberate strategy to localize the loss of sensitive information upon capture of an individual (Krebs 2002). Recent work has also considered how dependencies amongst multiple interdependent networks impacts robustness (Buldyrev et al. 2010; Helbing 2013). For example, energy stations on the power grid depend on communication nodes on the Internet for their control, and likewise, communication nodes depend on power stations for their electrical supply. Strong coupling between the two can trigger large failure cascades across both networks. When failures burst, high initial connectivity, while initially efficient, promotes the rapid loss of nodes.

The other extreme (highly economic topologies, such as stars and chains) are also rarely observed by themselves in real-world networks. These topologies yield much smaller failure bursts but also very low efficiency. The spectrum between these two ends can be parameterized by β, which measures the severity of catastrophic, correlated events of node loss (e.g. the black plague) that can occur during the evolution of a network. In other words, β measures how likely a node becomes targeted, attacked, or infected in the environment (resulting in its loss), and how quickly its neighboring nodes are also lost due to cascading effects. Ideally, topologies should adjust according to β; when $\beta = 0$, cliques are optimal. However, it is not clear what topologies are best for larger values of β, nor how to generate them algorithmically.

In short, we study the tension between node growth and node loss processes as applied to a wide variety of networks. We focus on two research problems here:

Problem 1 **Given:** a snapshot G of a time-evolving network, which operates in an environment with harshness β, **Find:** the value of β used to generate G.

In other words, we assume G evolved in a harsh environment with parameter β, and our goal is to reverse engineer this value of β using graph-theoretic measures (Sect. 3.1). As mentioned above, if the given network is clique-like, the ideal algorithm should guess that β is 0 or close to 0. To validate our algorithm, we correlate our predictions of β with true biological network robustness, which is measured experimentally and reflects the harshness of the environment (i.e. the probability that the cell will die following the loss of a node (gene) and downstream effects from this loss.)

Problem 2 Given: N, the number of nodes, and β, the environmental harshness, **Design:** a robust network G with N nodes.

Specifically, we want to design a graph G so that after bursty node loss with probability β, the residual graph (i.e. the graph after removing all lost nodes) is as highly connected as possible. To measure connectivity, we compare $\lambda_1(G)$ with $\lambda_1(G_{\text{residual}})$—i.e. the largest eigenvalue of the adjacency matrix of the original graph with the eigenvalue of the graph after removing all lost nodes (Sect. 4).

Overall, the contributions are the following:

- **A novel model**, MASSEXODUS: it is parsimonious, requiring only one parameter (β); it is realistic, capturing the behavior of biological networks; and it is useful, capable of forecasting the robustness of a network (i.e. the harshness of its environment in which bursty node loss follows from an initially targeted, attacked, or infected node).
- **A novel optimization problem** (namely, maximize the largest eigenvalue of the adjacency matrix *following* bursty node loss) with theoretical analysis for several classes of graphs.
- **An empirical evaluation** of the robustness of several real-world networks using this criteria.
- **A distributed, model-based algorithm** that is efficient and can be used to design robust networks according to the harshness of the environment.

2 Related work

The related work forms four sub-areas: (a) generative models for social networks; (b) generative models for biological networks; (c) studies of network robustness; and (d) epidemiological infection models. We are not aware of prior work that fully address our two research problems (how to predict the harshness of a network environment, and how to generate real networks under harsh environments).

Social network models. Most generative models of network formation assume unabated growth of nodes over time (Barabasi and Albert 1999; Broder et al. 2000; Fabrikant et al. 2003; Vazquez et al. 2003; Leskovec et al. 2005b, a; Siganos et al. 2006; Leskovec et al. 2008; Akoglu and Faloutsos 2009; Chakrabarti and Faloutsos 2012) or assume a constant number of nodes (Jin et al. 2001), again without node loss. One exception is the copying model of Kleinberg et al. (1999), which randomly removes individual nodes (webpages) based on their expected lifetime, though this model does not handle *bursty* node loss occurring alongside growth. Recently, Wu et al. (2013) studied node arrival and departure dynamics in social and collaboration networks and found that the likelihood of a user departing the network is closely related to the activity level of his neighbors (i.e. how often his friends send or receive content, or update their status). In contrast, our model considers departure more generally as being triggered independently by an external environment, as is the case in many scenarios discussed in the Introduction.

Biological network models. Biological network evolution is largely based on copying or *duplication* models, which have been shown to best reproduce topologies of

global protein interaction networks (Vazquez et al. 2003; Middendorf et al. 2005; Navlakha and Kingsford 2011). In the Vazquez et al. (2003) model, an existing ambassador node v spawns a topological duplicate u. Node u connects to node v with probability q_{con}. Then, for each common neighbor x of u and v, with probability q_{steal}, either the edge (u, x) or (v, x) is removed (chosen randomly), and with probability $1 - q_{\text{steal}}$ both edges to x are retained. Unlike previous copying models (Kleinberg et al. 1999), here the ambassador may also lose edges after duplication. This model is inspired by the general *duplication* principle in evolution (Vazquez et al. 2003), in which a duplicate gene is initially functionally redundant, but over time, the two genes diverge and specialize into sub-functions. The parameter q_{steal} directly controls how many shared neighbors of u and v are retained (i.e. whether it is full or partial duplication). Prior work has proposed parameters for this model that best mimic the topology of global protein interaction networks ($q_{\text{steal}} = 0.4$ and $q_{\text{con}} = 0.7$; (Navlakha and Kingsford 2011)). However, this model with fixed parameters cannot reproduce the range of topologies observed across robust to fragile biological subnetworks, as has been previously noted (Navlakha et al. 2014). Further, this model by itself ignores the effects of harsh environments (gene loss (Kitano 2004), environmental noise (Newman et al. 2006), or other such perturbations that can alter topology).

Network robustness. Most prior work on measuring network robustness have focused on optimization functions related to node or edge connectivity (Albert et al. 2000), natural connectivity (Chan et al. 2014), among others (Schneider et al. 2011; De Domenico et al. 2014), yet these measures can all be optimized using cliques, which, as argued in the Introduction, are not realistic nor robust when node loss is bursty.

Infection models. Many epidemiological models have been proposed to model cascading processes on graphs (Chakrabarti and Faloutsos 2012). Typically, in these models, all nodes are initially susceptible (S). Then, an initial node is infected (I) and this infected state is recursively passed on to susceptible neighbors (following edges in the graph) with probability β. Infected nodes transition to the recovered (R) state with probability δ, after which they are not susceptible to becoming infected again. In this paper, we use the SIR model because it is the most parsimonious model with all three states (susceptible, infected, and recovered).

2.1 Biology background

One of our main motivations is to model the structure and evolution of robust biological networks more realistically. Biological networks have evolved to optimize performance in the face of environment and signaling noise (Newman et al. 2006) and viral and bacterial attacks (Kitano and Oda 2006). These failures can be bursty and render some nodes and edges unusable. Similarly, at the evolutionary time-scale, gene mutation and loss is a common perturbation (Gu et al. 2003; Kitano 2004), the effects of which can also propagate through the network (Valencia and Pazos 2003; Albert 2007; Guell et al. 2012).

True biological robustness can be measured using genetic knock-out experiments, where a single gene (node) is experimentally removed from the genome (network). If this removal results in cell death, the gene is deemed essential or *fragile*, and otherwise

it is *robust*. Most genes [roughly 80 % in *S. cerevisiae* and *E. coli* (Giaever et al. 2002; Gerdes et al. 2003)] are robust, meaning the cell can survive the loss of these nodes (one at a time) and all downstream effects caused by their loss.

Interestingly, fragile genes primarily lie in portions of the network that are *least likely* to experience environmental noise, perturbations, and attack (Navlakha et al. 2014). Formally, global protein interaction networks are decomposable into modules or subnetworks consisting of all proteins (and their induced interactions) involved in a similar biological process (e.g. transcription, endosomal transport, etc.). Most information processing occurs within modules, and these modules lie physically embedded in various locations in the cell (e.g. transcription occurs internally, in the nucleus; endosome transport occurs close to the boundary membrane). External, membrane modules more exposed to harsh external environments are more robust; i.e. they contain few fragile genes. On the other hand, internal modules contain many fragile genes, however, these modules are also physically more protected from harsh environments. Thus, biological networks are shaped by their environments, yet to our knowledge, formal models of this process are lacking.

For both social and biological networks, node loss occurs alongside an underlying growth process. Next, we present a biologically-inspired model of network evolution that takes both of these processes into account.

3 Proposed model: MassExodus

We are given n, the number of nodes in the network, and β, the environmental harshness. For example, in biology, β may be a function of the network's distance from the external environment, as discussed above. On the Internet, secure or firewall-protected intranets will have lower β than public networks exposed to the external world.

We begin with two (active) nodes connected by an edge and with the remaining $n - 2$ nodes as *isolates*. A node is *active* if it has degree at least 1. In each epoch of the model, there is a node loss phase and a node growth phase, as described below.

For the loss phase, there are three states for nodes: susceptible, infected, or recovered (Easley and Kleinberg 2010). First, we select a random node u in the graph (active or isolated), and with probability β, infect u (all other nodes at this stage are susceptible). We then initiate an infection cascade (burst) starting from u. This burst follows a standard susceptible-infectious-recovered (SIR) epidemiological model with contact probability also β and recovery probability $\delta = 1$. This means each node has exactly one chance to pass the infection onto each of its susceptible neighbors in the graph (each transmission succeeding with probability β) before entering the recovered state. This process continues until there are no nodes remaining in the infected state. All nodes that were infected (and their incident edges) are then removed from the graph and isolated. In subsequent epochs, these nodes may be infected again, as well as any other non-infected node in the current epoch. This completes the node loss phase (Fig. 1).

For the growth phase, we select a random node from the isolated set (if there are any) and add it to the active graph using an existing generative growth model. In this paper, we focus primarily on the *duplication* model, in which an isolated node chooses

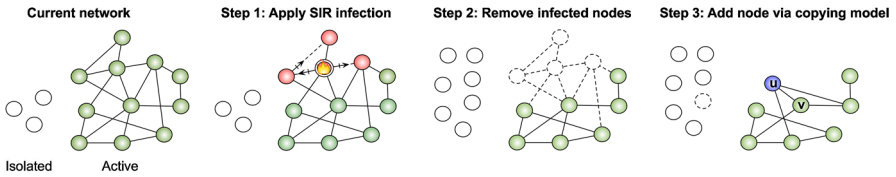

Fig. 1 Overview of the MASSEXODUS model. The current network consists of all *active* nodes with at least one incident edge and a set of *isolated* nodes. Step 1: A randomly chosen node is selected and, with probability β it is infected and triggers an SIR infection cascade (burst). Step 2: All infected nodes are removed from the network and placed into the isolated set. Step 3: A random node from the isolated set is added to the active network via the growth model

an active node, connects to it with probability q_{con}, and copies or steals some of its edges with probability q_{steal} (for full description, see Sect. 2). However, any growth model that proceeds by iteratively adding nodes to an existing graph can be employed within our framework (Barabasi and Albert 1999; Broder et al. 2000; Fabrikant et al. 2003; Vazquez et al. 2003; Leskovec et al. 2005a, b; Siganos et al. 2006; Leskovec et al. 2008; Akoglu and Faloutsos 2009; Chakrabarti and Faloutsos 2012). This completes the growth phase and one epoch of the model (Fig. 1). Both phases repeat for $4n$ epochs to provide sufficient time for the network to develop. Algorithm 1 shows pseudocode for the model.

Even if an active node is initially selected during the loss phase, this node only becomes infected (triggering an SIR burst) with probability β. Hence, there may be many growth-only epochs, where nodes are added to the active graph, before any nodes are removed.

All growth-only models are a special case of our model with $\beta = 0$. Our model is applicable to both undirected and directed networks.

3.1 Classifying network robustness

To address Problem 1, we want to predict the environmental harshness (β) of G under the MASSEXODUS model. In other words, we want to find:

$$\beta^* = \underset{\beta}{\text{argmax}} \Pr[G|\beta], \tag{1}$$

i.e. the value of β that maximizes the conditional probability that G evolved in environment β according to the model. To estimate this probability, we developed a regression framework to predict the likelihood that G evolved in environment β given just three topological features of G: the first (largest) eigenvalue of its adjacency matrix, the number of connected components in G, and the fraction of isolated nodes in G. These features were selected because they, together, can differentiate between 3 topological regimes we observed for evolving graphs under MASSEXODUS (before and after two phase transition values of β; Results, Sect. 5.3). These features were also highly correlated with biological fragility (Results, Figure 4A) amongst 10 features individually tested, and thus they represent a good benchmark for our model.

Algorithm 1 The MASSEXODUS model. The function *compute_SIR_burst(G,u,β)* returns the set of all infected nodes following infection of u using an SIR model. The function *run_duplication_model(u,v)* takes an isolated node u and connects it to an active node v in G using the duplication model.

Require: n (# of nodes) and β (environmental harshness)
#—Start with two nodes connected by an edge and $n - 2$ isolates—
G.add_edge(1,2)
Active ← set(1,2)

G.add_isolates([3 . . . n])
Isolated ← set(3, 4, . . . n)

#—Repeat until steady-state reached—
for $t \in [1, \ldots, 4n]$ **do**
  #—Select node, spread infection—
  if random() < β **then**
    u ← G.random_node()
    Infected ← compute_SIR_burst(G,u,β)

    #—Remove infected nodes—
    G.isolate_nodes(Infected)
    Active.remove(Infected)
    Isolated.add(Infected)
  end if

  #—Add node via duplication model—
  if |Isolated| > 0 **then**
    u ← Isolated.random_element()
    v ← Active.random_element()
    Isolated.remove(u)
    Active.add(u)
    G.run_duplication_model(u,v)
  end if

end for
return G

To facilitate fair comparison of feature values for different network sizes, we constructed a regression model $\mathcal{R}_n$ to classify networks with exactly n nodes. To train $\mathcal{R}_n$, we used the MASSEXODUS model to construct training networks using values of β ranging from 0.00 to 0.27 in step sizes of 0.03. For each β, we simulated the MASSEXODUS model and, following an initial burn-in period of $2n$ iterations to allow the network to develop, we observed ("sampled") networks with probability 0.10 in each epoch. Sampling ensures networks represent the range of topologies offered by β, which can oscillate since growth and loss probabilistically co-occur. For each sampled network, we extract a feature vector consisting of the three features mentioned above, which are then associated with the corresponding target value of β. Once all networks for each value of β are sampled, we built a nearest-neighbor regressor $\mathcal{R}_n$ to predict the value of β given the feature vector for G. The challenge of the regression task is to learn the range of topologies produced by each value of β and to differentiate between similar values of β.

Cross-validation was performed by comparing actual and predicted values of β; a prediction was considered correct only if the predicted value of β exactly matched the actual value (using a nearest-neighbor classifier; we also tried decision trees and

a Gaussian Naive Bayes model, but both performed worse than the nearest-neighbor model.) We used 50 different values of n ranging from 33 to 450. These 50 values corresponded to the exact values of n for the 50 real biological subnetworks, as discussed below.

3.2 Experimental setup and biological data description

We collected a global interaction network for *S. cerevisiae* (baker's yeast) by integrating protein-protein interactions and protein-DNA interactions from prior experimental studies (Chatr-Aryamontri et al. 2013; MacIsaac et al. 2006). The network consisted of 5,796 proteins and 79,988 undirected and unweighted interactions. We decomposed this global network into 50 subnetworks, with each subnetwork corresponding to a biological process. Annotations associating proteins to biological processes were taken from the Gene Ontology database (Ashburner et al. 2000). Each subnetwork consisted of all proteins associated with the process and their induced interactions. Because some biological processes may not be as inherently important for cell survival and growth as others (thereby affecting robustness requirements), we only selected processes that represented housekeeping processes without any of which the cell would likely die, as determined by an expert yeast biologist (Navlakha et al. 2014).

Of the 5,796 proteins, 19.4 % (1,122) were determined to be essential in normal growth conditions by Giaever et al. (2002) (i.e. removal of any of these single nodes caused the entire network to fail). Each subnetwork was then assigned a fragility score equal to the percentage of essential genes in the subnetwork. This percentage reflects the probability that the cell will die following the loss of a node (gene) and downstream effects from this loss.

4 Proposed robustness measure and analysis

Economy and efficiency are two common factors driving network evolution (Bullmore and Sporns 2012; Louf et al. 2013), but neither appears sufficient by itself to explain real-world topology. Economy is related to the amount of resources required to build a network, which in our case is simply the number of edges. Efficiency is the distance or number of hops separating two random nodes. For example, if economy were the only consideration, then most networks would have tree- or star-like topologies; for pure efficiency, cliques would naturally emerge. While some (latent) combination of these factors likely drives network evolution, these networks must also be robust to bursty node loss events potentially triggered by the environment. This leads to the following question: what topologies are most adequate for variably harsh environments?

To address Problem 2, we propose the following novel optimization problem that seeks highly efficient networks *following* bursty node loss. Formally, we seek:

$$G^* = \underset{G}{\operatorname{argmax}}\, \lambda_1(G_{\text{residual}}). \tag{2}$$

To overcome a bursty failure, a network should be designed such that the connectivity of its *residual network*, G_{residual}, is maximized. The residual network corresponds to

the remaining network after the removal of nodes lost in the SIR infection process. Connectivity is measured via the largest eigenvalue of the adjacency matrix (λ_1; Prakash et al. 2010). Comparing the eigenvalue before (G) and after (G_{residual}) node removal implicitly measures both the impact of the loss and the efficiency of the resulting topology. Finding G^* is challenging and depends on the parameter β used in the SIR process. If $\beta = 0$, cliques are optimal, but for higher values of β, cliques also enable swift spreading of the infection, leading to low residual connectivity.

Below, we analytically compute the expected largest eigenvalue of the adjacency matrix after bursty node loss for four classes of special graphs. We assume a random initial node is selected for the SIR process. We also assume G is connected (undirected) and n is the number of nodes in G.

Lemma 1 *For star graphs, the expected eigenvalue after loss is:* $\mathrm{E}[\lambda_1(G_{\text{residual}})] = (\frac{n-1}{n})(1-\beta)\sqrt{n}$.

Proof There are two cases corresponding to the two types of nodes that can be initially infected: the center hub and the satellites.

$$\text{Center} : \left\{ \frac{1}{n-1} \qquad 0 \right. \tag{3}$$

$$\text{Satellites} : \begin{cases} \frac{n-1}{n}(1-\beta) & \sqrt{n}-\epsilon \\ \frac{n-1}{n}(\beta) & 0 \end{cases} \tag{4}$$

The center hub is targeted with probability $1/(n-1)$ and its loss results in the isolation of every node, leaving a residual graph with eigenvalue 0. A satellite is targeted with probability $(n-1)/n$. If it does not pass the infection to the center hub, the residual graph has the same eigenvalue as the original ($\sqrt{n-1}$), minus some negligible constant, ϵ (Chung et al. 2003). If the satellite infects the center hub, the residual graph has eigenvalue 0. □

Lemma 2 *For line or chain graphs,* $\mathrm{E}[\lambda_1(G_{\text{residual}})] = 2$.

Proof The expected number of infected nodes along one side of the chain follows a geometric sum: $\sum_{n=0}^{\infty} \beta^n = 1 + \beta + \beta^2 + \beta^3 \cdots = \frac{1}{1-\beta} < \infty$, if $\beta < 1$. Similar analysis follows for the other side of the chain. As $n \to \infty$, the residual graph will thus also be a chain having eigenvalue approximately 2. □

For analysis of clique graphs, let the number of recovered nodes following an SIR process on a clique to be defined as: $R_\infty = 1 - S(0)\exp^{-R_0(R_\infty - R(0))}$, where $S(0) = \frac{n}{n-1}$, the proportion of initially susceptible individuals; $R(0) = 1/n$, the proportion of initially recovered individuals; and $R_0 = n \times \beta/\delta$.

Lemma 3 *For cliques,* $\mathrm{E}[\lambda_1(G_{\text{residual}})] = n - R_\infty - 1$.

Proof The residual graph will be a clique on all remaining uninfected nodes (of which they are $n - R_\infty$), thus having eigenvalue $n - R_\infty - 1$. □

Finally, to simplify analysis for Erdős–Rényi random graphs, we considered a SI model with two time ticks; i.e. the initial node is infected at time $t = 1$, passes the infection onto each neighbor independently with probability β at time $t = 2$, and then the infection propagation stops. This model represents a lower bound on the number of infected nodes versus using the SIR model. This model can clearly still infect hubs even if hubs are not directly targeted.

Lemma 4 *For an Erdős–Rényi random graph with parameter p,* $\mathrm{E}[\lambda_1(G_{\text{residual}})] = ((n-1-\beta(n-1)p-1)p+(1-p))$.

Proof The eigenvalue of an Erdős–Rényi random graph with n nodes and connection probability p is: $\lambda_1(G_{n,p}) = (n-1)p+(1-p)$ (Fredi and Komls 1981). The expected number of nodes in the residual graph after infection is $n' = n-1-\beta(n-1)p$, and this residual graph is also approximately an Erdős- -Rényi random graph. Thus, the expected eigenvalue after infection is $\lambda_1(G_{n',p})$. □

5 Results

We tackle the following four questions: First, can we predict the robustness of biological networks using the MassExodus model? Second, does the MassExodus model accurately reproduce the spectrum of topologies observed in non-biological networks (such as communication or social networks)? Third, does the MassExodus model exhibit any interesting phase transitions? And fourth, is robustness to bursty node loss a reasonable latent optimization function driving the evolution of real-world networks, and how can we design robust networks?

5.1 Predicting robustness of biological networks

Given a network G, can we predict how harsh the environment is in which G evolved? Or, what value of β was used to generate G using the MassExodus model?

We developed a regression model to predict β given just 3 topological features of G (Sect. 3.1). Predictions tested using tenfold cross-validation were highly accurate and increased with network size (Fig. 2a). For large networks ($n > 200$), accuracy reached 85–90 %, implying that the exact value of β was highly predictable among 10 possible classes of β values (we observed similar performance for $n = 1{,}000$). Performance was slightly worse for smaller networks (accuracy of 75–80 %) due to higher variability in the occurrence and impact of node loss events. This variability led to less separation of features in feature space across different values of β (Fig. 3). A baseline classifier that randomly predicts β would result in 10 % accuracy, thus our classifier yields a 7.88-fold improvement in performance, averaged across all n. Overall, β leaves its stamp on topology, and this stamp can be accurately captured with a simple predictor using three features.

Next, we show that MassExodus is a reasonable model for biological network evolution and can distinguish between robust and fragile biological subnetworks. Biological robustness is measured experimentally and reflects environmental harshness

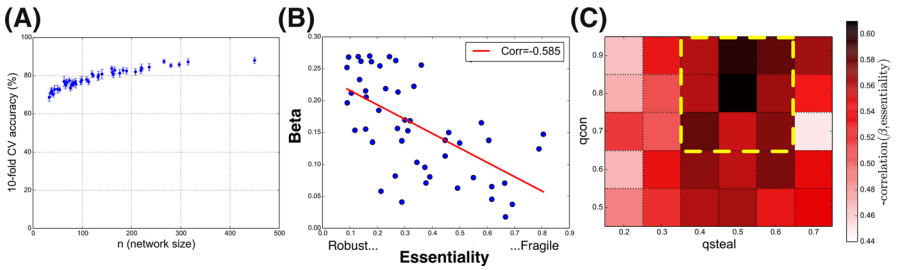

Fig. 2 Accurate prediction of network robustness and biological fragility. **a** Cross-validation accuracy of predicting β using a nearest neighbor classifier given networks generated using the MASSEXODUS model. **b** Correlation between actual biological fragility and predicted value of β (using MASSEXODUS and $q_{\text{steal}} = 0.4$, $q_{\text{con}} = 0.7$) for each biological network. Highly fragile networks exist in safe environments (low β), whereas highly robust networks operate in noisy environments (high β). **c** Heatmap of correlation coefficients between biological fragility and β using different parameters in the duplication model (q_{steal}, q_{con}). *Dotted yellow square indicates* range of highest correlation, matching literature-proposed parameters (Color figure online)

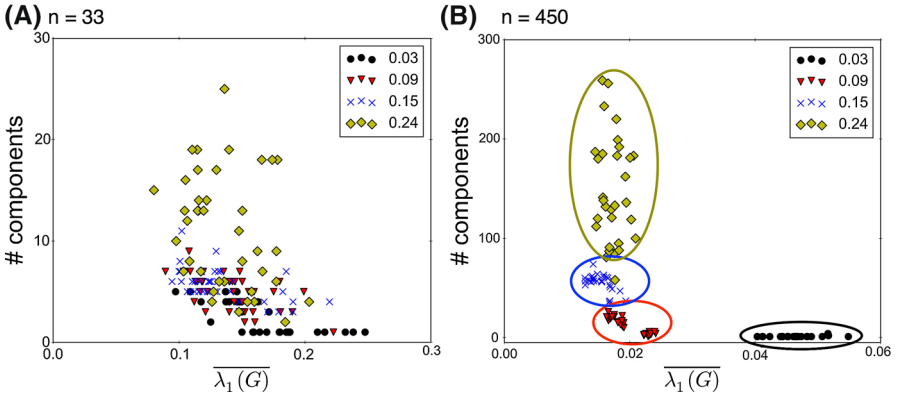

Fig. 3 Separability of topological features in feature space increases with network size. Each dot corresponds to a network generated using the MASSEXODUS model. *Colors* and *shape* indicate the value of β. **a** With small n, variability in the occurrence and impact of node loss events leads to less separability of features, making prediction of β difficult. **b** As n increases, networks generated using the same value of β are much more clustered. The two phase transitions, from *black* to *red/blue* to *yellow*, are also more apparent (Color figure online)

(Sect. 2.1). Here, we use β to measure environmental harshness and attempt to predict the value of β for each biological subnetwork. To do this, we considered each biological subnetwork G and built a regressor $\mathcal{R}_n$ with the same number of nodes n as G (Sect. 3.1). $\mathcal{R}_n$ was trained using MASSEXODUS networks and was then applied to topological features extracted from the *real* biological subnetwork. We repeated this for all 50 subnetworks and computed the correlation between the true, known fragility of the subnetwork and the predicted value of β.

Most fragile biological subnetworks were correctly predicted to have significantly lower values of β than robust subnetworks (Fig. 2b). This means that our model correctly inferred that fragile biological networks lie in *safe* environments, which is

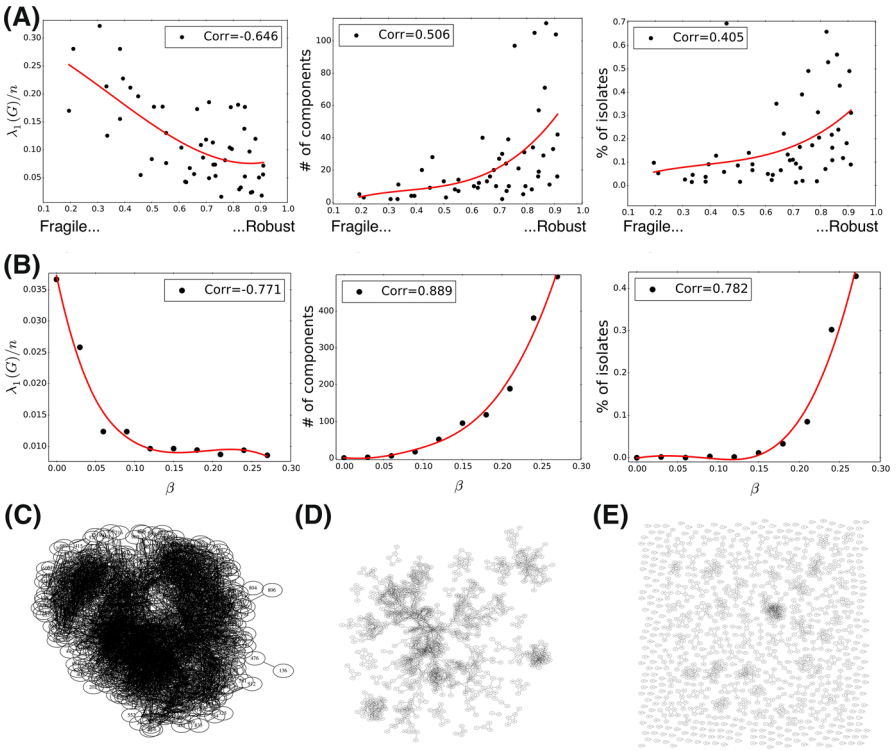

Fig. 4 The MASSEXODUS model reproduces the range of topologies observed within biological networks. **a** Comparison of how three topological features (normalized eigenvalue of the adjacency matrix, number of connected components, percentage of isolated nodes) vary as a function of 1-fragility for the 50 biological networks. **b** Similar comparison versus β for the MASSEXODUS model. Correlation coefficients between x and y axes are shown in the legend. *Red lines* depict spline interpolations of the data points, which show qualitatively similar behavior with the real data. **c**–**e** Three example networks generated using the MASSEXODUS model with $n = 1{,}000$ nodes and **c** $\beta = 0.03$, **d** $\beta = 0.09$, and **e** $\beta = 0.24$. Two phase transitions occur from one giant connected component → several medium-sized components → mostly isolates (Color figure online)

indicative of their physical (internal) positioning in the cell. On the other hand, robust networks were predicted to have evolved in a highly noisy and deleterious environment (large values of β), as their external positioning also suggests. Interestingly, 5 of the 50 subnetworks we analyzed had β values that did not agree with their biological function (DNA-dependent transcription elongation, chromatin organization, and transcription from RNA polymerase II promoter, snoRNA processing and ribosomal subunit export from nucleus). Some of these may be explained as errors in the initial harshness annotated to the module (e.g. even though ribosomal export from the nucleus was labeled as a module existing in a more noisy environment, all activity of proteins in this module occurs within the cell, and these proteins have little to no contact with external signals). Further investigation is required to determine if the other modules represent true biological anomalies or potential errors in the model.

The high correspondence between β and biological fragility suggests that the MASSEXODUS model matches the spectrum of topologies observed across robust and fragile biological networks (Fig. 4a, b). Robust biological subnetworks tended to have smaller eigenvalues with more connected components and isolated nodes. As mentioned above, such sparseness may be a deliberate evolutionary strategy to localize the spread of noisy environmental signals within harsh environments (high β). On the other hand, fragile biological networks had higher connectivity, and fewer connected components and isolated nodes, as this enables more efficient signaling in safe environments (low β). The MASSEXODUS model also yields similar qualitative trends when comparing topological features with environmental harshness, β (Fig. 4b).

Comparison and advantages over prior work. Prior analyses suggests that fragile biological networks lie in safe environments (low β, where robustness is less important), and as a result, foster highly connected topologies to enhance efficiency (Navlakha et al. 2014). On the other hand, robust networks lie in harsher environments (high β) and have sparser topologies to isolate the influence of bursty perturbations. Thus, the cell has adjusted the topology of networks based on their susceptibility to spreading perturbations and mutational effects, and our model accurately captures these differences in an evolutionary growth model.

How dependent is regression accuracy on parameters of the underlying growth model? The experiments above used fixed values of $q_{\text{steal}} = 0.4$ and $q_{\text{con}} = 0.7$, while varying β. We repeated the experiment using values for $q_{\text{steal}} \in [0.2 - 0.7]$ and $q_{\text{con}} \in [0.5 - 0.9]$. For each parameter pair, we built a regression model and computed the correlation between biological fragility and predicted value of β for the 50 subnetworks. Remarkably, the highest correlations observed use parameters very similar to the optimal parameters proposed in the literature for this model ($q_{\text{steal}} = 0.5 \pm 0.1, q_{\text{con}} = 0.8 \pm 0.1$ here, versus $q_{\text{steal}} = 0.4, q_{\text{con}} = 0.7$ in the literature (Navlakha and Kingsford 2011); Fig. 2c). The decay in correlation for very low or high values of q_{steal} suggests that too little or too much divergence following duplication confers topologies that are not realistic, which also agrees with prior analysis of interaction conservation between protein homologs using protein structure and sequence-based analysis (Ispolatov et al. 2005; Pereira-Leal et al. 2007). This further suggests that the MASSEXODUS model may be capturing actual dynamics of biological evolution.

Finally, the spectrum of topologies observed in real biological networks may also be mimicked by employing different growth model parameters for different environments, without resorting to any node loss events. For example, for fragile (internal) networks, low values of q_{steal} could be used to generate denser topologies, whereas for robust networks, higher values of q_{steal} can generate sparser networks. To test this, we built a regression model trained using networks grown with different values of q_{steal} and computed the correlation between biological fragility and the predicted value of q_{steal} using the same three features as above. The correlations were similar: 0.628 versus 0.609 with MASSEXODUS and $q_{\text{steal}} = 0.5, q_{\text{con}} = 0.8$. There are, however, two downsides to the previous approach. First, it does not model node loss caused by mutation, gene loss, and other such perturbations known to affect network connectivity (Gu et al. 2003; Kitano 2004; Albert 2007; Guell et al. 2012). Second, more importantly, it requires that the network somehow know *a-priori* about its environment

Table 1 Summary statistics of real-world networks used

Network	# nodes	# edges
Yeast protein interactions	5,796	79,988
C. elegans brain	279	2,287
Europe road	1,174	1,417
US airports	1,574	17,215
USA Powergrid	13,579	37,448
Gnutella-0808	6,301	20,777
Gnutella-0825	22,687	54,705
World airports	2,939	30,501
Hamsterer social	2,426	16,631

All graphs are undirected

so growth can be guided by pre-selecting the appropriate value of q_{steal}. In practice, such an oracle is not present. For example, in biology, duplication parameters would need to somehow adjust on-the-fly to environmental harshness, and it is not clear what the molecular mechanism of this could be. Likewise, conditions in monitoring environments for sensor networks may change drastically over time and without warning and thus requires on-the-fly adaptation. For our model, the environment (β) implicitly drives the network into a robust topology without any knowledge or change in the underlying growth mechanism.

Analysis of other model variants. We also tested two additional node loss models. For the first, we fixed the node loss phase to occur every 10 epochs (instead of the loss phase occurring in every epoch with probability β) starting from a random node (isolated or active). For the second, we used an SI model (Chakrabarti and Faloutsos 2012), instead of an SIR model, but with limited number of rounds. In each round, all infected nodes pass the infection to each neighbor with probability β. There is no recovery, but there are only 5 rounds, after which the infection propagation stops, and all infected nodes are removed from the graph and isolated. Both of these models resulted in similar qualitative behavior, and with similar phase transitions (Sect. 5.3). Results are omitted for brevity.

5.2 Generality of the MassExodus model applied to non-biological networks

Can the MassExodus model reproduce topological features observed in non-biological networks? To test this, we collected several real-world networks (transport, communication, and social networks; Table 1) and compared their degree distributions, distributions of top eigenvalues, and average shortest path lengths versus networks generated by the MassExodus model.

We observed strong agreement between the real and model-generated topologies (Fig. 5a). We present these comparisons under a range of β values for different networks to explore the impact of different levels of node loss. MassExodus was used with the duplication model (Vazquez et al. 2003) or the forest fire model (Leskovec et al. 2005b) during growth, and model parameters were chosen to produce networks with exactly the same number of nodes as the real-networks (static snapshots) with roughly

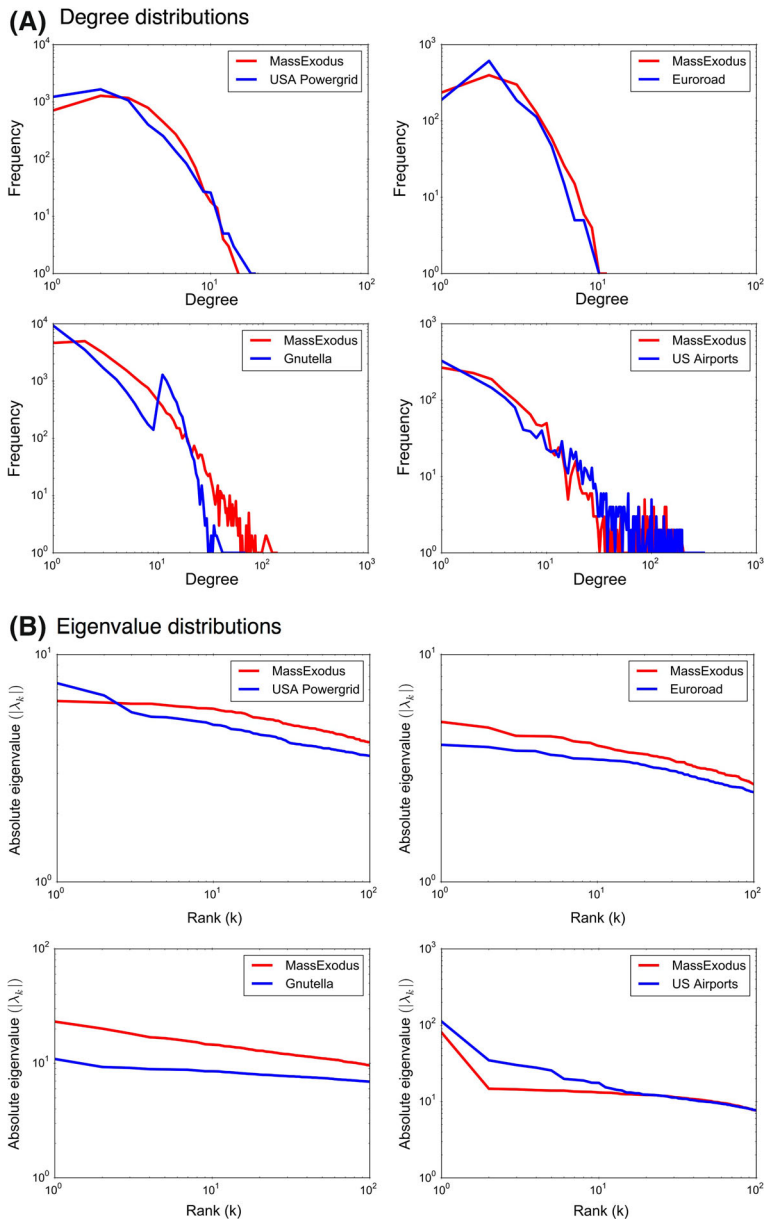

Fig. 5 Close correspondence between degree and eigenvalue distributions of real-world and MASSEXODUS networks. **a** Degree distributions and **b** Distributions of the top 100 real eigenvalues of the adjacency matrix. Parameters used were as follows: USA Powergrid (duplication model with $q_{\text{steal}} = 0.7, q_{\text{con}} = 1.0$ and $\beta = 0.09$), European road network (duplication model with $q_{\text{steal}} = 0.8, q_{\text{con}} = 0.9$ and $\beta = 0.05$), Gnutella-0825 P2P network (forest fire model with $p = 0.40$ and $\beta = 0.025$), and US Airport network (forest fire model with $p = 0.50$ and $\beta = 0.009$)

the same number of edges (see Fig. 5 legend for parameters used). Interestingly, our model can produce both power-law (Gnutella peer-to-peer network and USA airport network) and non-power-law degree distributions (the two geometric networks, USA Powergrid and European road network). Our model also closely matched the distribution of the top 100 real eigenvalues for all four networks (Fig. 5b). Finally, real and model-generated networks also had similar average shortest path lengths between nodes (14.9 vs. 18.4 for the European road network, comparing MASSEXODUS and real; 3.3 vs. 3.1 for US airports; 17.2 vs. 19.0 for USA Powergrid; and 6.6 vs. 5.5 for Gnutella-0825 network).

5.3 Phase transitions

Interestingly, as β increased, we observed two phase transitions in network connectivity. The first transition occurred between $\beta = 0.06$ and $\beta = 0.09$, where a single giant connected component was split into several medium-sized components (Fig. 4c, d). The second transition occurred between $\beta = 0.21$ and $\beta = 0.24$, when the smaller components were further split into mostly isolated nodes (Fig. 4d, e). While the exact β values for the phase transitions depend on the parameters of the the underlying growth model (in our case, q_{steal} and q_{con}), the transitions were still observable across many other parameter settings.

5.4 Observations from evaluating robustness of real-world networks

The close correspondence between three general topologies properties of our model and real-world networks suggests that robustness to bursty node losses may be an implicit factor driving network evolution. It also suggests that networks today should ideally be built to withstand more than just single node loss. For example, road construction can result in the closure of several, nearby roads in an area, as opposed to just a single road. For air travel, if bad weather halts travel to Chicago, then other nearby cities will also likely be affected, as well as smaller cities that route through Chicago. Next we use our new optimization criterion (Sect. 4) to analyze the robustness of real-world networks to bursty node loss.

We consider two types of node failures (Albert et al. 2000):

1. **Random failure**: where a single random node is infected and the infection spreads using an SIR process.
2. **Adversarial failure**: defined as the worst case connectivity following SIR infection of the 100 highest degree nodes.

Figure 6 shows λ_1 of the original graph G (x-axis) versus λ_1 of the residual graph (y-axis). Robust networks should lie close to the 45° line—indicating minimal disruption following failure—while also maximizing the residual eigenvalue (i.e. the connectivity following infection). This latter requirement is essential since otherwise, sparse networks with small original eigenvalues would be optimal.

Observation 1 (Robustness to random, bursty failures) *Real-world networks are robust to random bursty failures (Fig. 6a).*

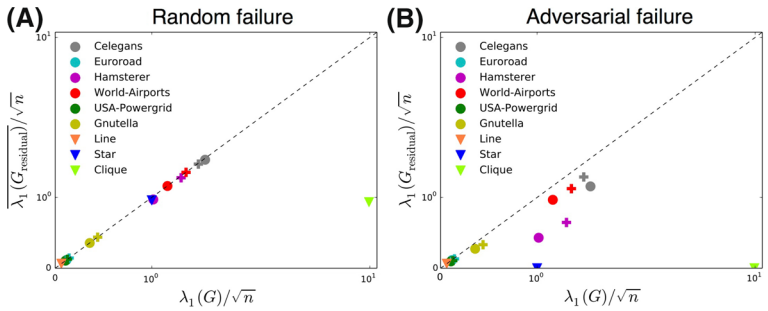

Fig. 6 Real-world and MASSEXODUS networks are robust to random and adversarial bursty failure. Eigenvalue of the original network (x-axis) versus eigenvalue of the residual graph after removing nodes lost due to bursty failure (y-axis)—under **a** random failure and **b** adversarial failure. Real-world networks are shown as *circles*, and MASSEXODUS networks are shown as '+' and are color-coded to match the real-network they try to mimic. For each network pair, the same $\beta < 0.05$ was used. *Inverted triangles* show special-case graphs (*chain*, *star*, *clique*), all with poor performance

Most networks showed less than 1 % reduction in residual eigenvalue, with the largest drop occurring for the Hamsterer social network (4 %). This suggests that several real-world networks have evolved topologies that remain fault tolerant under random bursty failure.

Observation 2 (Mixed behavior to adversarial failures) *Several, but not all, networks were significantly disrupted following adversarial failure (Fig. 6b).*

The Hamsterer social network showed the largest drop in the residual eigenvalue following adversarial node loss (57.6 %), whereas the *C. elegans* brain network, the Gnutella peer-to-peer network, and the World Airports travel network showed a smaller drop (25.6, 23.5, and 16.9 %, respectively). This difference could be because the latter three networks implicitly evolved for overcoming faults and perturbations, whereas Hamsterer likely did not evolve for resilience (the Hamsterer social network no longer exists today). It also highlights the fact that, although adversarial attacks are rare, they can potentially lead to catastrophic events. Two of the geometric networks (USA Powergrid and the European road network) were largely unaffected by even adversarial failure, likely due to their mesh-like, non-power-law topology, which isolates failures to mostly local regions. These networks, however, also had the smallest original eigenvalue of all real-world networks.

Next, we use the MASSEXODUS model to design robust networks.

Observation 3 (Comparing MASSEXODUS and real-world topologies) *The* MASSEXODUS *networks closely mimic the behavior of the real-world networks under both random and adversarial attack.*

In Fig. 6, real-world networks are shown as circles, and MASSEXODUS networks are shown as '+' and are color-coded to match the real-network with the same number of nodes and roughly the same number of edges. The close correspondence in the original eigenvalue and the residual eigenvalue following bursty node loss of both MASSEXODUS and real-world networks suggests: (a) that our robustness measure may

also be a realistic (latent) optimization criteria driving network evolution; and (b) that our distributed algorithm (based on MASSEXODUS) can be used to construct networks tailored to operate within harsh environments; e.g. for distributed sensor or wireless networks in hazardous conditions.

Observation 4 (Unrealistic behavior of textbook graphs) *The three special graphs (lines, stars, and cliques) exhibited very different behavior.*

Lines were robust to both random and adversarial failure but had a very small residual eigenvalue (2). Stars were robust to random failure—when satellites were targeted and did not propagate the infection to the center node—but were completely disconnected following adversarial attack of the center node. Cliques were not robust to either attack, but had a very large original eigenvalue.

Observation 5 (Comparison to prior robustness measures)

Our robustness measure is in stark contrast to prior measures of robustness (Related Work), for which cliques are optimal for both types of attacks. This suggests that either real-world networks are very far from optimal, or that these measures, alone, are not realistic optimization functions driving network evolution. In contrast, our measure leads to more realistic graphs and combines both robustness and efficiency, which are two common criteria used to evaluate networks.

6 Conclusions

We studied the dynamics of node growth and node loss processes on evolving networks. In particular, we contributed the following:

- **A novel model**, MASSEXODUS, the first that handles bursty node loss, while also generalizing other growth-only models. MASSEXODUS can guess the environmental harshness (biological fragility), and it can mimic the spectrum of topologies observed in many real-world networks.
- **A novel optimization problem** (namely, maximize the largest eigenvalue of the adjacency matrix *following* bursty loss) and evidence that this latent function drives network evolution.
- **An empirical evaluation** of how several real-world networks fare to this new robustness criteria, and a new distributed algorithm to design networks in harsh and adversarial environments.

Questions for future work include: First, we empirically observed a double phase transition in connectivity as β increased; closer analytics of these transitions and their potential relation to connectivity in random graphs would be interesting to pursue. Second, we measured robustness of the residual graph using λ_1, but recent work has proposed natural connectivity (Chan et al. 2014) as an alternative measure of robustness that characterizes alternative paths and the topology of all connected components. It would be interesting to see if natural connectivity is also relevant to biological networks and how to optimize it within our model. Third, we assumed a synchronous order of events but did not consider cases where multiple nodes join and leave the network at the same time, or cases where nodes have non-uniform loss probabilities.

7 Reproducibility

An implementation of our model is available at: www.snl.salk.edu/~navlakha/mass_exodus/.

Acknowledgments This material is based upon work supported by the National Institutes of Health award no. F32-MH099784 to S.N; by the National Science Foundation under Grant No. IIS-1217559, by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and by a Google Focused Research Award to C.F; and by grants from the McDonnell Foundation programme on Studying Complex Systems and from the US National Science Foundation award nos. DBI-0965316 and DBI-1356505 to Z.B.-J. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2013) Systemic risk and stability in financial networks. NBER Working Papers 18727, National Bureau of Economic Research Inc. http://ideas.repec.org/p/nbr/nberwo/18727.html

Akoglu L, Faloutsos C (2009) Rtg: a recursive realistic graph generator using random typing. Data Min Knowl Discov 19(2):194–209. doi:10.1007/s10618-009-0140-7

Albert R (2007) Network inference, analysis, and modeling in systems biology. Plant Cell 19(11):3327–3338

Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382

Albert R, Albert I, Nakarado GL (2004) Structural vulnerability of the North American power grid. Phys Rev E Stat Nonlinear Soft Matter Phys 69(2 Pt 2):025,103

Alippi C, Galperti C (2008) An adaptive system for optimal solar energy harvesting in wireless sensor network nodes. IEEE Trans Circuits Syst I 55(6):1742–1750. doi:10.1109/TCSI.2008.922023

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. In: Proceedings of the 9th international world wide web conference on computer networks: the international journal of computer and telecommunications netowrking, North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, pp 309–320. http://dl.acm.org/citation.cfm?id=347319.346290

Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464(7291):1025–1028

Bullmore E, Sporns O (2012) The economy of brain network organization. Nat Rev Neurosci 13(5):336–349

Carle J, Simplot-Ryl D (2004) Energy-efficient area monitoring for sensor networks. Computer 37(2):40–46

Chakrabarti D, Faloutsos C (2012) Graph mining: laws, tools, and case studies. Synthesis lectures on data mining and knowledge discovery. Morgan & Claypool Publishers, San Rafael, CA

Chan H, Akoglu L, Tong H (2014) Make it or break it: manipulating robustness in large netwotks. In: SIAM International conference on data mining (SDM)

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41(Database issue):D816–D823

Cho A (2013) Computer science. Network science at center of surveillance dispute. Science 340(6138):1272

Chung F, Lu L, Vu V (2003) Eigenvalues of random power law graphs. Ann Comb 7:21–33

De Domenico M, Sole-Ribalta A, Gomez S, Arenas A (2014) Navigability of interconnected networks under random failures. Proc Natl Acad Sci USA 111(23):8351–8356

Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, New York, NY

Fabrikant A, Luthra A, Maneva E, Papadimitriou CH, Shenker S (2003) On a network creation game. In: Proceedings of the twenty-second annual symposium on principles of distributed computing, ACM, New York, NY, PODC '03, pp 347–351. doi:10.1145/872035.872088

Fredi Z, Komls J (1981) The eigenvalues of random symmetric matrices. Combinatorica 1(3):233–241. doi:10.1007/BF02579329

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol 185(19):5673–5684

Giaever G, Chu AM, Ni L, Connelly C et al (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418(6896):387–391

Gidalevitz T, Prahlad V, Morimoto RI (2011) The stress of protein misfolding: from single cells to multicellular organisms. Cold Spring Harb Perspect Biol. doi:10.1101/cshperspect.a009704

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421(6918):63–66

Guell O, Sagues F, Serrano MA (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621

Haldane AG, May RM (2011) Systemic risk in banking ecosystems. Nature 469(7330):351–355

Helbing D (2013) Globally networked risks and how to respond. Nature 497(7447):51–59

Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. Nucleic Acids Res 33(11):3629–3635

Jin EM, Girvan M, Newman MEJ (2001) Structure of growing social networks. Phys Rev E 64(046):132. doi:10.1103/PhysRevE.64.046132

Kitano H (2004) Biological robustness. Nat Rev Genet 5(11):826–837

Kitano H, Oda K (2006) Robustness trade-offs and host-microbial symbiosis in the immune system. Mol Syst Biol 2(2006):0022

Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. In: Proceedings of the 5th annual international conference on computing and combinatorics, Springer, Berlin, COCOON'99, pp 1–17. http://dl.acm.org/citation.cfm?id=1765751.1765753

Krebs V (2002) Mapping networks of terrorist cells. CONNECTIONS 24(3):43–52

Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C (2005a) Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In: Proceedings of the 9th European conference on principles and practice of knowledge discovery in databases, Springer, Berlin, PKDD'05, pp 133–145. doi:10.1007/11564126_17

Leskovec J, Kleinberg J, Faloutsos C (2005b) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th international conference on knowledge discovery and data mining, pp 177–187. doi:10.1145/1081870.1081893

Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks. In: Proceedings of the 14th international conference on knowledge discovery and data mining, pp 462–470. doi:10.1145/1401890.1401948

Louf R, Jensen P, Barthelemy M (2013) Emergence of hierarchy in cost-driven growth of spatial networks. Proc Natl Acad Sci USA 110(22):8824–8829

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. BMC Bioinf 7:113

Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. Proc Natl Acad Sci USA 102(9):3192–3197

Moore D, Shannon C, Brown J (2002) Code-Red: a case study on the spread and victims of an Internet worm. SIGCOMM/USENIX Internet Measurement Workshop. Marseille, France, pp 273–284

Moore D, Shannon C, Voelker G, Savage S (2003) Internet quarantine: requirements for containing self-propagating code. In: Proc. of the IEEE Intl. Conf. on Computer and Communications, vol 3, pp 1901–1910 vol. 3, doi:10.1109/INFCOM.2003.1209212

Navlakha S, Kingsford C (2011) Network archaeology: uncovering ancient networks from present-day interactions. PLoS Comput Biol 7(4):e1001,119

Navlakha S, He X, Faloutsos C, Bar-Joseph Z (2014) Topological properties of robust biological and computational networks. J R Soc Interface 11(96):20140,283

Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 441(7095):840–846

Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. Genome Biol 8(4):R51

Prakash B, Tong H, Valler N, Faloutsos M, Faloutsos C (2010) Virus propagation on time-varying networks: theory and immunization algorithms. In: Proceedings of the European conference on machine learning and knowledge discovery in databases, Springer, Berlin, pp 99–114. http://dl.acm.org/citation.cfm?id=1889788.1889796

Schneider CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ (2011) Mitigation of malicious attacks on networks. Proc Natl Acad Sci USA 108(10):3838–3841

Siganos G, Tauro SL, Faloutsos M (2006) Jellyfish: a conceptual model for the as internet topology. J Commun Netw 8(3):339–350. doi:10.1109/JCN.2006.6182774

Sole RV, Rosas-Casals M, Corominas-Murtra B, Valverde S (2008) Robustness of the European power grids under intentional attack. Phys Rev E Stat Nonlinear Soft Matter Phys 77(2 Pt 2):026,102

Valencia A, Pazos F (2003) Prediction of protein–protein interactions from evolutionary information. Methods Biochem Anal 44:411–426

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. Complexus 1(1):38–44

Wu S, Das Sarma A, Fabrikant A, Lattanzi S, Tomkins A (2013) Arrival and departure dynamics in social networks. In: Proceedings of the sixth ACM international conference on web search and data mining, ACM, New York, NY, WSDM '13, pp 233–242. doi:10.1145/2433396.2433425